Combinatorial Optimization in Biology

Principal Investigator:

Madhav V. Marathe, CIC–3
7–8010, marathe@lanl.gov

Co-Investigators:

Allon G. Percus, CIC–3/T–CNLS
5–7435, percus@lanl.gov

David C. Torney, T–10
7–9452, dct@lanl.gov

Technical Category: Computer Science and Software Engineering (CSSE)

FY 2000 Funding Requested: $140K

Abstract:

Quantitative technologies have, in recent years, taken on enormous importance for molecular biology. Some of the most fundamental problems facing biologists today, such as sequencing the Human Genome, rely extensively on the use of bioinformatics and computational techniques. Large gaps remain to be filled, however. Little attempt has been made to address some important optimization problems arising in a biologically-motivated setting, often because these problems have been thought to be computationally hard.

Surprisingly, many of them are not. Particularly for problems relying on a geometric formulation, the theoretical machinery often exists for finding algorithms that solve them in polynomial time. Our goal is, based on these theoretical insights, to develop such algorithms so that they can readily be implemented.

In the case of the Human Genome Project, we have already found that simple models of DNA sequencing procedures lead to polynomial-time algorithms that can improve sequencing efficiency substantially. We intend to extend these algorithms to be able to accommodate additional experimental desiderata. This will enable full-scale automation of DNA sequencing and the elimination of a labor-intensive bottleneck that currently retards the process. Another optimization problem we intend to address in this context is that of pooling, or group testing. Pooling is the technique of choice for identifying a few "positive" clones from a large collection consisting mainly of "negatives". While designing efficient yet accurate pools appears at first sight to be computationally hard, experience with sequencing algorithms suggests to us that useful variants on the problem may in fact be solvable in polynomial time. We intend to develop algorithms for constructing implementable pooling designs that will robustly yield the desired identifications, using a minimum number of pools.

Scientific and Technical Impact

Biological problems are now at the scale where sophisticated computer algorithms are essential. This fact, and its realization by the biological community, has been among the principal factors responsible for the extremely rapid growth in biotechnologies recently. In algorithmic theory, however, supply has not kept up with demand. The theoretical foundations of combinatorial optimization, developed by computer scientists over decades, have been slow to migrate over to this new arena. The algorithmic work we propose is intended to fill some fairly major gaps, providing approaches to be implemented at the forefront of biological research. In addition to establishing the credentials of optimization theory in applications to the life sciences, it will impact some of the most high-profile laboratory programs, most notably the Human Genome Project and follow-on projects such as that in Functional Genomics.

Los Alamos, as part of the DOE's Joint Genome Institute, is in a prime position to derive benefit from improved algorithms in bioinformatics. Our research will contribute significantly to this end. For example, our work on optimized DNA sequencing algorithms could, based on preliminary results, lead to an efficiency gain of 25% over current methods. The need for novel DNA sequencing technologies has been articulated repeatedly, at all levels of the project: achieving the current goal of sequencing the entire Human Genome by 2003 will, according to the latest DOE estimates, "require a two- to threefold improvement" over current technologies [1]. The development of pooling design algorithms for DNA clones is equally vital, as this will result in more efficient screening experiments. Our Genome Center at LANL has developed and implemented the most advanced pooling techniques in the world. Further efficiency gains, however, could save millions of dollars and free up valuable resources. Well constructed sets of pools can be implemented for a given clone library, and used over and over again in future experiments. The impact of improved efficiency in these areas translates into a national and worldwide imperative, as the scientific and medical importance of understanding the Human Genome is monumental. Even in its early stages, the Human Genome Project was recognized by Larry Deaven and Robert Moyzis, writing in *Los Alamos Science*, as "one of the most exciting and challenging research programs in the history of science" [2].

Background

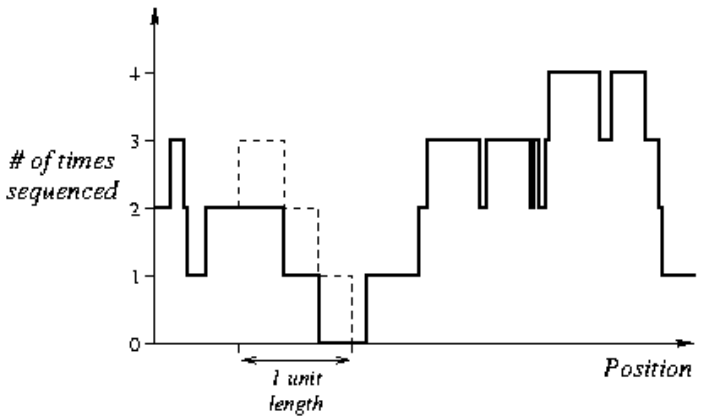
Over the past fifteen years, combinatorial optimization problems have become increasingly interesting to scientists in other fields. The Traveling Salesman Problem, for instance, has been studied extensively by statistical physicists, who observed that the algorithmic theory developed for that problem could be of great use in related physical systems (spin glasses are a noted example [3]). And conversely, notions arising from theoretical physics have contributed to better algorithms and a deeper and more accurate understanding of the solutions to combinatorial optimization problems. (Good examples of this are simulated annealing [4] and finite-size scaling [5].)

The links between optimization algorithms and molecular biology, however, have not on whole received the same degree of attention. And yet, the need is no less pressing. Modern biological research often involves processing a staggering amount of data. The Human Genome Project is perhaps the most dramatic example of this: the total genetic code contained in human DNA involves 24 chromosomes, each made up of approximately 10^8 nucleotides. The identity of each of these nucleotides must be determined. And then, when the sequence is known, in order for it to be biologically useful it must be organized into extensively tested DNA libraries representing individual genes. Clearly, if this project is to be completed in an acceptable amount of time and if

human/material resources are to be used wisely, every effort must be made to optimize the process and to reduce inefficiency to an absolute minimum. Let us now outline two areas that we believe lie at the intersection of good science and efficiency.

The sequencing problem. Sequencing a strand of DNA simply means determining the order in which the nucleotides, or bases, occur on the strand. DNA sequencing experiments [6] are typically performed on *clones*, copies of the chromosomal DNA that are of the order of 10^5 bases in length. These experiments consist of two phases: *shotgun sequencing* and *walking*. Shotgun sequencing may be thought of as a stochastic process, where many short subintervals at random locations on a DNA strand are sequenced. A fixed cost is associated with each shotgun-sequenced subinterval. Walking is a deterministic finishing process, where regions insufficiently covered in the shotgun process may be sequenced. Unlike shotgun sequencing, the locations for walking are under good experimental control, but the procedure is also much more labor-intensive and time-consuming. A higher cost is associated with each subinterval sequenced in this way. Note that both procedures sequence *discrete* subintervals, which may all be taken to have equal (unit) length.

Current standards in the Human Genome Project [7] require in practice that every position on the DNA strand be sequenced at least 3 times, to insure minimum reliability of results. Any moderate amount of shotgun sequencing will tend to leave *some* regions of the DNA strand insufficiently sequenced according to this criterion; once the locations of these regions have been established, they must be finished by walking.[1] The following figure depicts a possible "coverage profile" on part of the clone, showing the initial shotgun layout as well as a (unit-length) walking location:



This presents the experimenter with two optimization problems. The first is deciding the right balance of how much (inexpensive, but uncontrolled) shotgun sequencing to perform, versus how much (expensive and labor-intensive, but controlled) walking to perform, in order to achieve finished sequence according to the required criterion. The second is, given a certain amount of shotgun sequencing, how precisely to perform the walks. Since shotgun positions are random, the insufficiently sequenced regions are generally of non-unit length, as seen in the figure above. Walking experiments, however, sequence one unit length at a time. It is therefore a non-trivial problem to decide *where* to place the walks, in order to minimize the expense of meeting the criterion.

[1] The current algorithms assume that the positions of sequenced intervals are known, which is not always the case, a point that will be redressed in the current project.

Solving these optimization problems are important, as they are responsible for a major labor bottleneck in the Human Genome Project. To our knowledge, they have barely been addressed in the literature. One possible reason for this is that, at first sight, it appears that they might be *NP-hard*, *i.e.*, no algorithm exists that can solve the problem to optimality in polynomial time. To our own surprise, our preliminary research has convinced us that this is false [9]. The problem of placing the walks can be solved by a linear-time greedy algorithm, and even a generalized version, involving economy of multiple walks, can be solved in polynomial time by a dynamic programming approach (with good linear-time approximations). We have reason to believe that when further experimental realities are incorporated, such as the double-stranded nature of DNA and a probability distribution for sequenced interval lengths, the problem will remain computationally tractable. This is of course a vital concern when the DNA clones being modeled contain, typically, 10^5 bases.

As this research is new, no directly relevant publications by others can be cited here. However, two helpful sources of related information are Ref. [1] and Ref. [10].

Pooling designs. Once a large amount of sequenced genome data has been made available to the scientific community, the ultimate goal will be to guide the research process in a more functional direction, trying to gain a synthetic view of gene function. This requires high-quality gene libraries, which in turn requires extensive testing and screening of the DNA clones used in the libraries. Pooling of clones and, potentially, of proteins is essential for these objectives. In addition to locating clones containing a particular DNA sequence, appropriate pools could identify the subtle peculiarities of gene collections predisposing to common maladies, such as hypertension or various types of tumors.

Thus, pools of (say) cloned DNA are screened rather than the clones individually. Although considerable overlap between pools is necessary in order to identify accurately the clones that test positive in any given screening experiment, judicious construction of the pools can still result in large efficiency gains. For example, in the Life Sciences division at LANL in FY98, 376 pools were constructed using a total of 220,000 clones, each pool containing approximately 5,000 clones in all. This was done in such a way that almost no information was lost in the process. An experimental task was therefore reduced from screening 220,000 clones individually to screening just 376 separate pools.

Given the ensemble of pool screening results, we use a "decoding" procedure for finding the probability that each clone is actually positive. The certainty with which clones can be identified as positive, together with the number and size of the pools used, is a measure of the pooling design's quality. This is what must be maximized. Efficient pooling is important, because once a pooling design of DNA clones is implemented it can of course be used in an unlimited number of future experiments. All of these experiments will benefit from the gain in efficiency. Thus, well-designed pools are an invaluable resource for clone libraries. Designing optimally efficient pooling strategies is, however, a challenging problem in combinatorial design theory. Existing work on this has been limited to pools of small size; for pools containing thousands of clones, the problem is not obviously tractable. In fact, efficient implementable designs exist only at LANL.

The optimal combinatorial design to use in the absence of experimental constraints is likely to be what is known as a t-design [11]. We have, so far, used heuristics to find good approximations to t-designs under realistic constraints for constructing the pools. We believe, however, that there is a good chance of developing algorithms to improve upon this. Moreover, we have no firm reason for expecting all versions of the constrained optimization problem to be NP-hard; we would not be surprised if near-optimal solutions (with performance guarantees) or even optimal solutions could be found in polynomial time.

Ref. [11] is representative of the state-of-the-art in the field of combinatorial designs for group

testing.

Considerable interest has been shown by the Life Sciences division in both of the two areas mentioned above. Los Alamos intends to perform large-scale DNA sequencing (5 million bases per year) in support of Joint Genome Institute objectives. Thus, algorithms that lend themselves to automation of high-throughput sequencing are essential. These would readily be implemented in the laboratory at LANL. Furthermore, some of our heuristically optimized pooling designs have already been implemented experimentally, within the context of Joint Genome Institute programs. Norman Doggett (LS-3) has indicated his strong interest in having LS division collaborate further, on the experimental end, on all of the projects discussed here.

Also, it should be noted that research performed under LDRD/DR grant #98806, *Probabilistic and Combinatorial Analysis of Biological Systems* (PI: David Torney), was especially helpful in leading us to develop our present proposal.

Research Objectives and Goals

The main goals of our research are as follows:

- Introduce new algorithmic approaches for combinatorial optimization in biologically-motivated problems, such as those arising in the Human Genome Project. Show that useful models can be formulated and that optimal solutions or good near-optimal solutions can be found in polynomial time.
- Elaborate upon our preliminary models, in order to take full account of biological and experimental realities, as well as multiple finishing criteria.
- Use our sequencing algorithms to enable automation of high-throughput sequencing at Joint Genome Institute facilities, including here at LANL.
- Create polynomial-time algorithms that rapidly generate high-efficiency implementable DNA clone pools, for use with future clone libraries.

R&D Approach

First of all, we plan on using the current formulation of our sequencing models as a basis for a more sophisticated, realistic and useful approach. This means employing a dynamic programming method that we have developed for optimizing the placement of walking intervals on the DNA clones under generalized cost conditions. The method involves a decomposition of walking strategies into subproblems that use shorter clone segments. Recursing from smaller subproblems to larger ones, and bootstrapping on subproblem solutions that have already been found in the procedure, we obtain an optimal solution in $O(n^3)$ time, where n is the clone length measured in units of walking intervals. We will expand upon this method to incorporate the full double-stranded address problem, where error-reduction standards pose the additional requirement that *both* strands be sequenced everywhere by either shotgun or walking intervals. Understanding the algorithmic modifications imposed by these further constraints (which should be trivial) will allow us to design realistic sequencing algorithms that could then be implemented at LANL. This will provide us with valuable real-world data on realized efficiency gains; we intend to analyze simulated data as well, to test our models.

Second of all, we would like to enlarge these models to include scheduling aspects. This means improving efficiency by using the fact that several sequencing machines could be running at once, allowing, for instance, walks on different parts of the clone to be performed in parallel. (At Los Alamos, we have access to a Mermaid primer synthesizer, which generates the primers for 96 walking experiments in one batch.) This introduces a new set of costs, more accurately reflecting the reality of labor costs versus equipment costs. As Dr. Alejandro Schaeffer of the National Center for Bioinformatics at the National Library of Medicine, NIH, has pointed out in discussions with us, accurate models and optimization strategies involving this are lacking, and are urgently needed for planning purposes when large-scale sequencing equipment is to be purchased.

Finally, we intend to develop an algorithmic framework for pooling strategies, in much the same way as we have done for sequencing. Using this framework, we expect to be able to explore the question of NP-hardness for optimal pooling design. Although we are presently inclined to doubt that the problem is indeed NP-hard, we should ultimately be in a position to determine the problem's complexity — either by finding a polynomial-time algorithm, or by proving that this is impossible. One immediate aim is ϵ-approximation algorithms, so that we can formulate (and justify) the best *practical* means of constructing such pools. We expect to have ample opportunities to implement these ideas on ongoing work of the LANL Center for Human Genome Studies.

We are fortunate to have, in our present team, researchers with strong credentials in all of these areas. Madhav Marathe has made extensive contributions to the areas of computational complexity theory, performing pioneering work on the notion of approximation algorithms for multi-criteria optimization problems. Allon Percus has worked both in mathematical biology and, for some years, on combinatorial optimization strategies motivated by statistical physics; in his research on the stochastic traveling salesman problem, he produced asymptotic estimates for expected tour lengths that are the most precise to date. David Torney has worked on innumerable aspects of sequencing and pooling strategies, having published seminal papers on pooling designs with error detection.

Expected Scientific and Technical Results

The results that we expect to obtain will include, among others:

- *A polynomial-time algorithm for optimizing the placement of walks on the full double-stranded sequencing problem.* Although this complicates the geometric formulation, we believe that it does not increase the computational complexity of the problem, and that a dynamic programming method will still yield optimal solutions in $O(n^3)$ time (with good linear-time approximations).

- *A sequencing model that superimposes scheduling constraints, based on the degree of allowed parallelism in the walking procedure.* This will introduce parameters representing specific aspects of sequencing costs, taking into account issues such as staffing and equipment overhead, rather than rolling them all into one or two single cost parameters.

- *A polynomial-time algorithm for optimal or near-optimal pooling designs under experimental pool construction constraints, i.e., an algorithm yielding designs that are minimally different from t-designs.* We also plan on designing optimized "templates", for simplifying the experimental construction of pool designs.

Funding Breakout

Funding at or near the FY2000 level is requested for 3 years. The breakdown for the first year is expected to be as follows:

$65K = 1/2 TSM or postdoc (Allon Percus)
$32K = 1/4 TSM (Madhav Marathe)
$32K = 1/4 TSM (David Torney)
$10K = part-time GRA and/or visitor funds

Key LANL staff participating in the project are expected to be:

Madhav V. Marathe (CIC–3)
Mark O. Mundt (CIC–12)
Allon G. Percus (CIC–3/T–CNLS)
David C. Torney (T–10)

External collaborations are planned with the National Center for Bioinformatics at the National Library of Medicine, NIH. Expected scientific visitors from this institution include Dr. Eva Czabarka, who has collaborated with us on preliminary work.

Possible Specialist Reviewers

Internal reviewers:

- Norman Doggett, LS–3 and Joint Genome Institute
  5–4007, doggett@lanl.gov
- Alan Perelson, T–10
  7–6829, asp@lanl.gov

External reviewers:

- Richard Karp, Dept. of Computer Science, University of Washington
  E-mail: karp@cs.washington.edu
- Ron Shamir, Dept. of Computer Science, Tel-Aviv University
  E-mail: shamir@math.tau.ac.il

References

[1] See `http://www.ornl.gov/hgmis/research.html`.

[2] L. L. Deaven and R. K. Moyzis, "The Los Alamos center for human genome studies", *Los Alamos Science* **20** (1992) 6–7, available at `http://lib-www.lanl.gov/pubs/number20.htm`.

[3] S. Kirkpatrick and G. Toulouse, "Configuration space analysis of travelling salesman problem", *J. Phys. France* **46** (1985) 1277–1292.

[4] S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi, "Optimization by simulated annealing", *Science* **220** (1983) 671–680; V. Černy, "Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm", *J. Optimization Theory Appl.* **45** (1985) 41–51.

[5] A. G. Percus and O. C. Martin, "Finite size and dimensional dependence in the Euclidean traveling salesman problem", *Phys. Rev. Lett.* **76** (1996) 1188–1191.

[6] F. Sanger, S. Nicklen and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors", *Proc. Nat. Acad. Sci. USA* **74** (1977) 5463–5467.

[7] Human Genome Program, U.S. Department of Energy, "JGI and 'Bermuda-quality' sequence", *Human Genome News* **9**:3 (1998) 7. Available at: `http://www.ornl.gov/hgmis/publicat/hgn/v9n3/07bermud.html`. See also: `http://www.gene.ucl.ac.uk/hugo/bermuda2.htm`.

[8] F. Alizadeh, R. M. Karp, L. A. Newberg and D. K. Weisser, "Physical mapping of chromosomes: a combinatorial problem in mathematical biology", *Algorithmica* **13** (1995) 52–76.

[9] A. G. Percus and D. C. Torney, "Greedy algorithms for optimized DNA sequencing", *Proc. of SODA '99* S955–S956.

[10] R. M. Karp and R. Shamir, "Algorithms for optical mapping", *Proc. RECOMB 98* 117–124. Available at: `http://www.math.tau.ac.il/~shamir/papers.html`.

[11] C. J. Colbourn and J. H. Dinitz, *The CRC Handbook of Combinatorial Designs* (CRC Press, Boca Raton, 1996), pp. 564–565.

MADHAV V. MARATHE

Research Interests: - Modeling, Simulation and Analysis of Large Scale Systems, High Performance Computing, Design and Analysis (theoretical and experimental) of Algorithms, Complexity Theory, Combinatorial Optimization, Data Mining, Mobile Computing

Education:
- University at Albany, SUNY: Ph.D. August 1994, Computer Science
- Indian Institute of Technology, Madras, India: B.Tech. August 1989, Computer Science

Work Experience:
- June 1996 - Present: Technical Staff Member, LANL
- July 1998 - Present: Adjunct Assistant Professor, Department of Computer Science, University of New Mexico
- August 1994 - June 1996: Postdoctoral Research Associate, LANL

Reviewer:
- Associate Editor: Journal of Computing and Information
- Reviewer: SIAM J. on Computing, Theoretical Computer Science, Operations Research, International Journal on Foundations of Computer Science, INFORMS J. Computing, Networks, Information Processing Letters, ACM-SIAM Symposium on Discrete Algorithms, FST & TCS, Symposium on Parallel and Distributed Computing

Selected Publications:

1) "Loop Transformations for Performance and Message Latency Hiding in Parallel Object-Oriented Frameworks," F. Bassetti, K. Davis, M.V. Marathe and D. Quinlan, in: Proc. International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 98), Las Vegas Nevada, July 1998.
2) "Theory of Periodically Specified Problems: Complexity and Approximability," M.V. Marathe, H.B. Hunt III, D.J. Rosenkrantz and R.E. Stearns, in: Proc. 13th IEEE Conference on Computational Complexity, Buffalo, NY, June 1998.
3) "Map Labeling Problems," S. Doddi, M.V. Marathe, A. Mirzaian, B. Moret, and B. Zhu in: Proc. 8th ACM-SIAM Symposium on Discrete Algorithms (SODA), San Francisco, CA, pp. 148-157, January 1997.
4) "Complexity of Hierarchically and 1-Dimensional Periodically Specified Problems I: Hardness Results," M.V. Marathe, H.B. Hunt III, R.E. Stearns and V. Radhakrishnan, in: AMS-DIMACS Volume Series on Discrete Mathematics and Theoretical Computer Science: Workshop on Satisfiability Problem: Theory and Application, Vol 35, pp. 225-259, November 1996.
5) "Approximation Algorithms for the Minimum Satisfiability Problem," M.V. Marathe and S.S. Ravi, in: Information Processing Letters, (IPL), Vol. 58, No. 1, pp. 23-29, April 1996.
6) "Approximation Schemes for PSPACE-Complete Problems for Succinct Specifications," M.V. Marathe, H.B. Hunt III, R.E. Stearns and V. Radhakrishnan, in: Proc. 26th Annual ACM Symposium on the Theory of Computing (STOC), pp. 468-478, May 1994.

ALLON PERCUS

Contact Information: Tel: 5-7435, E-mail percus@lanl.gov
WWW: http://www.lanl.gov/home/percus

Present Position: *Postdoctoral Research Associate* at Los Alamos National Laboratory, jointly affiliated with the Computer Research and Applications group (CIC-3) and the Center for Nonlinear Studies (CNLS)

Education:
- *Université Paris-Sud*: PhD in statistical physics
  September 1997
- *Ecole Normale Supérieure*, Paris: DEA (MS equivalent) in theoretical physics
  September 1994
- *Harvard University*: BA cum laude (physics)
  June 1992

Selected Publications:
- S. Boettcher and A.G. Percus, "Extremal optimization: methods derived from co-evolution", *Proceedings of GECCO–99*, to appear.
- A.G. Percus and O.C. Martin, "The stochastic traveling salesman problem: Finite size scaling and the cavity prediction", *Journal of Statistical Physics* 94 (1999), to appear.
- A.G. Percus and D.C. Torney, "Greedy algorithms for optimized DNA sequencing", *Proceedings of the 10th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '99)*, S955–S956.
- A.G. Percus and O.C. Martin, "Finite size scaling universalities of kth-nearest neighbor distances on closed manifolds", *Advances in Applied Mathematics* 21 (1998) 424–436.
- N.J. Cerf, J. Boutet de Monvel, O. Bohigas, O.C. Martin and A.G. Percus, "The random link approximation for the Euclidean traveling salesman problem", *Journal de Physique I* 7 (1997) 117–136.
- A.G. Percus and O.C. Martin, "Finite size and dimensional dependence of the Euclidean traveling salesman problem", *Physical Review Letters* 76 (1996) 1188–1191.
- H.M. Lacker and A. Percus, "How do ovarian follicles interact? A many-body problem with unusual symmetry and symmetry-breaking properties", *Journal of Statistical Physics* 63 (1991) 1133–1161.

LANL Activities:
- Co-organizer of *Frontiers of Combinatorics* workshop, Summer 1998
- Member of *Reseach Library Advisory Board*

Invited Presentations:
- "Greedy algorithms for optimized DNA sequencing", invited talk at National Center of Biotechnology Information, National Library of Medicine, Bethesda, MD, 15 January 1999.
- "The stochastic traveling salesman problem and the random link approximation", Physics Dept. Colloquium, Emory University, Atlanta, 13 November 1998.

DAVID COLTON TORNEY

PRESENT EMPLOYMENT

Theoretical Biology and Biophysics Group, Los Alamos National Laboratory. Staff member: 1987–present. Member of the Los Alamos Center for Human Genome Studies; responsible for designing large-scale clone mapping experiments of human chromosomes and pooling experiments.

EDUCATION

B.A. in Biology, cum laude, Harvard University, 1977
Ph.D. in Biophysics, minor in Mathematics, Stanford University, 1982
(Thesis: A Contribution to the Formulation of Diffusion-Limited Chemical Reaction)
M.D., Stanford University, 1984

PROFESSIONAL EXPERIENCE

Post-Doctoral Fellow, Theoretical Biology and Biophysics Group, LANL, June 1984–Sept. 1987
Staff Member, Theoretical Biology and Biophysics Group, LANL, Oct. 1987–present

SELECTED PUBLICATIONS/ARTICLES

1. W.J. Bruno, E.Knill, D.J. Balding, D.C. Bruce, N.A. Doggett, W.W. Sawhill, R. L. Stallings, C.C. Whittaker and D.C. Torney. "Efficient Pooling Designs for Library Screening". Genomics 26:21-30 (1995).
2. G. Xie, R. Lobb, W.J. Bruno, D.C. Torney and J.M. Gatewood. "Single-Base Sequencing and Similarity Comparisions." Genomics, 30:445-449 (1995).
3. N.A. Doggett, L.A. Goodwin, J.G. Tesmer, L.J. Meincke, D.C. Bruce, L.M. Clark, Mr. Altherr, A.A. Ford, H.-C Chi, B.L. Marrone, J.L. Longmire, S.A. Lane, S.A. Whitmore, M.G. Lowenstein, R.D. Sutherland, M.O. Mundt, E.H. Knill, W.J. Bruno, C.A. Macken, D.C. Torney, J.R. Wu, J. Griffith, G.R. Sutherland, L.L. Deaven, D.F. Callen, and R.K. Moyzis. "An Integrated Physical Map of Human Chromosome 16." Nature 377:335S-3655S (1995).
4. D.J. Balding, W.J. Bruno, E. Knill and D.C. Torney. "A comparative survey of non-adaptive pooling designs", in "Genetic mapping and DNA sequencing", IMA Volumes in Mathematics and its Applications, T.P Speed & M.S. Waterman eds., Springer Verlag, pp. 133-154 (1996).
5. D. J. Balding and D. C. Torney. "Optimal Pooling Designs with Error Detection." J. Comb. Theory A 74:131-140 (1996).
6. E. Knill, A. Schliep, and D.C. Torney. "Interpretation of Pooling Experiments Using the Markov Chain Monte Carlo Method." J. Comp. Biology 3:395-406 (1996).
7. D.J. Balding and D.C. Torney. "The Design of Pooling Experiments for Screening a Clone Map." Fungal Genetics and Biology 21:302-307 (1997).
8. W.J. Bruno, F. Sun, and D.C. Torney. "Optimizing Non-Adaptive Group Tests for Objects with Heterogeneous Priors." SIAM J. of Applied Mathematics 58:1043-1059 (1998).
9. E. Knill, W.J. Bruno and D.C. Torney. "Non-Adaptive Group Testing in the Presence of Errors." Discrete Applied Mathematics 88:261-290 (1998).
10. D.C Torney. "Sets Pooling Desings. Annals of Combinatoriecs (in press) (1999).
11. M. A. Chateauneuf, C.J. Colbourn, E.Lamken, D.R. Kreher and D.C. Torney. "Lattice Square Row-Column Front, and Union Jack Designs". Annals of Combinatoriecs (in press) (1999).
12. W.J. Bruno, G-C. Rota and D.C. Torney, "Probabilitiy Set Functions:. Advances in Applied Mathematics, In press (1999).
13. D.C. Torney, C. C. Whittaker, and G.Xie. "The Stationary Statistical Properties of Human Coding DNA Sequences." J. of Molec. Biol. 286:1461-1469 (1999).